Development of a Framework Dealing with Partial Data Unavailability and Unstructuredness to Support Post-Market Surveillance

Yijun Ren
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Italy
yijun.ren@polimi.it

Enrico Gianluca Caiani
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Italy
enrico.caiani@polimi.it

***Abstract*— Under the European Union Medical Device (MD) Regulation 2017/745, expert panel's decision on providing a scientific opinion on the Clinical Evaluation Assessment Report for high-risk MD is required, as part of the conformity assessment procedure. To this aim, the perceived risk of similar MDs already on the market, based on the European Medical Device Nomenclature (EMDN), could help. To generate such information, we propose a generalized framework to automatically collect and display in an aggregated way the publicly available safety notices (SNs), even when characterized by partial unstructuredness and incompleteness. This novel approach was tested on the Dutch data, consisting of 3618 SNs from 2015 to 2022, retrieved from the official government website by Web scraping. After the identification of named entities, the best match MD was searched within the Italian and Portuguese datasets of devices using Natural Language Processing techniques. Algorithm performance was tested on potentially equal SNs (472) published by both the Dutch and Italian authorities: assignment of the same EMDN code at level 1 was present in 454 out of 472 (96.19%) SNs, at level 2 in 447 (94.70%) SNs, at level 3 in 433 (91.74%) SNs. The proposed approach was able to cope with public data unavailability and incompleteness, thus providing structured data with appropriate EMDN usable for aggregation and safety signal detection.**

Keywords—Medical device regulation, Incomplete data, Natural language processing, Post-market surveillance, Safety signal detection, Web scraping

I. Introduction

The European Union (EU) Medical Device Regulation 2017/745, effective since May 26th, 2021, aims at creating a robust, transparent, and sustainable regulatory framework by bringing EU legislation in line with technological advances [1]. One of its novelties is the concept of post-market surveillance plan (PMS) for each medical device (MD), including the process of collecting, recording, and investigating complaints and reports related to MDs, to favor early detection of possible safety problems [2]. Based on the intended purpose and inherent risk of MDs, they are classified into: Class I (low-risk), Class IIa (medium-risk), Class IIb and Class III (high-risk), with a clinical investigation mandatory for new high-risk devices. In addition, for class III MDs, the responsible notified body will be obliged to request additional scrutiny as Clinical Evaluation Consultation Procedure, of its Clinical Evaluation Assessment Report (CEAR) by an Expert Panel (EP) composed of clinical experts, where the EP shall decide within 21 days whether to provide a scientific opinion on the CEAR, based also on the perceived risk of similar MDs on the market.

In the context of the EU project CORE-MD (Coordinating Research and Evidence for Medical Devices [3]), we proposed a framework to automatically collect and display in an aggregated way publicly available official content regarding MD alerts and recalls [4]. This was initially tested on data published by the Italian Ministry of Health, consisting of the safety notices (SNs) in the Italian Market, also exploiting the use of the European Medical Device Nomenclature (EMDN), characterized by an alphanumeric structure established in a seven-level hierarchical tree, mutated from the Italian classification system [5]. In this work, we propose a generalization of the proposed framework to be used for those countries where the published data are not completely available or are not properly structured. This new approach will be tested on relevant data from the Netherlands.

II. Materials and Methods

The proposed methodological steps are summarized in the framework depicted in Fig. 1. Each step will be described in the following sections.

A. Data Sources

Field SNs publicly available on the official website of the Health and Youth Care Inspectorate (IGJ) [6], part of the Dutch Ministry of Health, Welfare and Sports, were used. The IGJ supervises healthcare and youth care services in the Netherlands, as well as the market for medicines and MDs. Manufacturers are required to track MDs once they are placed on the market and to inform the IGJ about possible risks by a Field SN, then published on the official website.

Also, the list of all devices available in a specific market and the relevant EMDN codes are required for the process to work [4]. As such a list does not exist for the Netherlands, the lists of all MDs on the Italian and Portuguese markets, respectively available by the Italian Ministry of Health [7] and by Infarmed [8], were utilized, to generate the Dataset of Devices (DoD). Such lists were selected as they provided for each MD the corresponding EMDN code. These lists are continuously updated and included 1,703,175 devices for the Italian market (last update January 8, 2023) and 1,346,977 devices for the Portuguese market (last update January 10, 2023).

This research was conducted thanks to the funding received from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 965246, project "Coordinating Research and Evidence for Medical Devices (CORE-MD)".

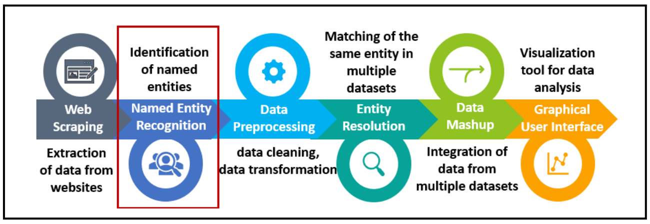

Figure 1. Methodological framework where the novelty step with respect to the previous implementation [4] is highlighted in red.

TABLE I. RELEVANT INFORMATION IN THE ITALIAN AND PORTUGUESE LISTS OF MDS ON THE MARKET

Country	Original Variable Name	Modified Variable Name in the DoD
Italy	Manufacturer/assembler	Manufacturer
	Commercial name	Device
	CND (the Italian “Classificazione Nazionale dei Dispositivi medici”) [a]	EMDN
Portugal	Manufacturer	Manufacturer
	Model	Device
	Brand	
	NPDM (the Portuguese Nomenclature for Medical Devices) [a]	EMDN

a. corresponds to the EMDN.

The variables reported in Table I, describing the original variables in the Italian and Portuguese lists and their modified names to be merged in the DoD (and used afterward) are reported. For the Portuguese dataset, the field *Device* was created by merging the fields *Brand* and *Model*, if not equal, due to possible data inconsistency and imprecision found in the list.

B. Web Scraping

Web scraping of the IGJ webpage [6], which reports all Dutch SNs, was performed using Python [9], [10], to retrieve the content of 3618 SNs dating from 2015 to 12/09/2022, thus constituting the local Dataset of Notices (DoN). For each notice, the following fields from the HTML text were retrieved:

- *Title*: webpage title that contains information about the manufacturer and MD.
- *Summary*: text that summarizes the SN.
- *Manufacturer*: name of the manufacturer.
- *Device*: name of MD.
- *Date*: last update date.

The fields *Manufacturer* and *Device* were present only in 2727 (75.37%) SNs; for the remaining 891 SNs, an additional step to identify such information from the field *Title* present in every SN, described in the next paragraph, was required.

C. Named Entity Recognition

Named Entity Recognition (NER), a subtask of Natural Language Processing (NLP), aims to identify named entities in a text [11]. There are four main streams of techniques applied in NER: rule-based [12], unsupervised learning [13], feature-based supervised learning [14], and deep learning (DL)-based approaches [15]. In recent years, the last ones become dominant and achieved state-of-the-art results thanks to their ability to discover complex and intricate features from data, their effectiveness in learning, and the possibility of training in an end-to-end approach [16]. In particular, the Transformer, the first sequence transduction model proposed in [17] and based solely on self-attention, has shown impressive performance on different NLP tasks. Self-attention can relate different positions of a single sequence to compute a representation of the sequence itself, thus capturing “long-term” dependencies within the sequence. Another key element contributing toward the development of these models is that of pre-training on a large corpus and then fine-tuning to the target task with a small labeled dataset to address the lack of adequate large dataset by transferring the knowledge of pre-trained model to a new one [18].

To allow generalization of the framework applicability to different countries, as SNs are not guaranteed to be reported in English, models trained from multiple languages were preferred over those trained only from English texts. Accordingly, the pre-trained transformer model XLM-RoBERTa [19] was used: XLM-RoBERTa is a transformer-based multilingual masked language model pre-trained on text in 100 languages with a massive dataset, that has superseded previous transformers, such as mBERT [18]. For the computational aspects, the Python open-source library Transformers was used [20], [21]. XLM-RoBERTa was then finetuned on the NER task using the 2727 SNs with identified information. These SNs were split into the test set (273 SNs, 10%), the training set (2208 SNs, 81%), and the validation set (246 SNs, 9%). For data preparation, the title of each SN was first tokenized by SentencePiece [22] and then annotated with ORG (organization) and MD tags in the “inside-outside-beginning 2” format [23]. The model was trained for 10 epochs, and the model checkpoint with the highest micro-averaging F1-score [24] evaluated on the validation set was selected and used to predict such information for the remaining 891 (24.63%) SNs without it.

D. Data Preprocessing

Real-world data tend to be incomplete, inconsistent, and missing. Data preprocessing techniques, aiming at improving data efficiency, have an essential role in data analysis [25]. As all the variables in DoD and DoN are string variables, except *Date* in the DoN, proper text preprocessing techniques were required to address poor quality issues: text transliteration [26], removal of brackets, extra spaces and punctuations, and lowercasing. For *Manufacturer* in both datasets, company name parsing was also applied to provide cleaned names by stripping away terms referring to organization type [27].

E. Entity Resolution

Entity Resolution (ER) describes the problem of extracting, matching, and resolving entity mentions in structured and unstructured data, thus identifying items in multiple data that refer to the same real-world entity [28]. In this work, the identification of the same MD in both DoD and DoN was necessary to associate each MD in the latter with the corresponding EMDN reported in the former. The problem was tackled in two phases to avoid unnecessary computational complexity and higher computational time: 1) identify similar manufacturers in both datasets; 2) identify similar devices among those records with similar manufacturers.

In the first phase, the field *Manufacturer* in both datasets was mapped into term frequency-inverse document frequency (TF-IDF) representations, then the cosine similarity (COS) was computed [29], considering as similar a score equal to or higher than 0.90.

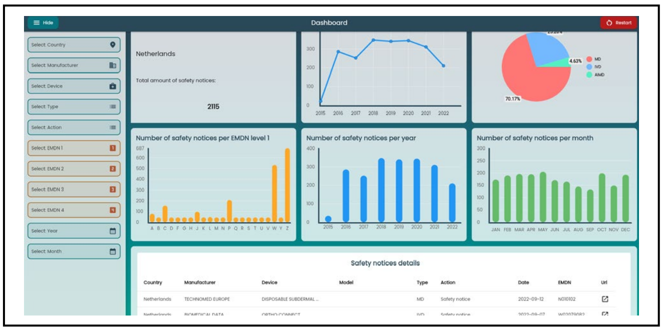

Figure 2. Graphical user interface to represent and query aggregated information.

In the second phase, the field *Device* was compared using approximate string matching [30]. In particular, fuzzy string matching was implemented [31], thus providing a similarity score from 0 to 100 based on the Levenshtein distance (LD). The minimum similarity threshold was set to 60, and the pair with the highest similarity score was considered as correspondent. As a result, if matching was found between MDs in two datasets, the relevant EMDN was retrieved. Due to the possible data inconsistencies encountered in the Portuguese DoD, the Italian one was used with priority for the ER.

F. Data Mashup and Graphical User Interface

A data mashup merges different homogenous or heterogenous data sources into a unique content page [32]. In our scenario, information within the DoN and the DoD was combined to provide a more complete dataset. Then, a Graphical User Interface (GUI) was created using Flutter, built and open-sourced by Google [33], to display the SNs in the DoN that were matched to devices in the DoD based on the user-defined queries. Fig. 2 shows the developed GUI: the left panel shows fields for query selection (i.e., country, manufacturer, device, type, action, EMDN levels, year, month), while the right panel shows the resulting SNs, using different automatically adapting visualization tools. Data trends and patterns over time are shown as line chart, while pie chart displays the distribution of SNs for different MD types, and the histograms are used to visualize the distribution of SNs for different EMDN at level 1, as well as the distribution of SNs along years or months. Additionally, the scrolling table at the bottom summarizes the resulting SNs, also providing the link to the original website.

III. Model Validation

To measure the model's performance, different tests were conducted on the NER and the ER subtasks. For the former one, the precision, recall, and F1-score were reported for each class evaluated on the test set, as well as the micro-averaging, the macro-averaging, and the weighted-averaging F1 [24].

The validation of the ER task was not straightforward due to the lack of a gold standard, as the list of all MDs on the Dutch market is unavailable, and thus there was no verifiable linkage between the DoN and the DoD. To address this issue, a different solution was adopted, based on the previous results successfully assigning EMDN codes to the Italian SNs [4]: if the same SN was found in both the Italian and the Dutch markets, then the assigned EMDN code at different levels could be used as the gold standard to determine the model's performance.

To ensure the correspondence of the two SNs, their attached pdf files were analyzed to verify identify of the reason for reporting, as the same manufacturer may report different issues regarding the same MD. As not all these pdf files are in English, only numeric characters were extracted from the file and used to calculate the similarity by means of the Jaccard similarity index (JSI) [34], thus avoiding the uncertainty related to the translation process. As a result, the same SNs published in both markets could be identified having a JSI=1, and the assigned EMDN code up to the fourth level was compared with the ER results, as a measure of model performance.

Realizing that even if the two SNs are identical, the calculated JSI may result in a value ≤ 1 due to possible discrepancies in the dates, additional numbers, or pages, in the respective pdf files due to the misalignment of reporting procedures, SNs that are potentially the same were also identified by loosening this condition and including additional conditions for manufacturer and device similarities, calculated using COS and LD, respectively. To this aim, several possible combinations of thresholds for JSI (from 0.5 to 1, step 0.1), COS (from 0.5 to 1, step 0.1), and LD (from 50 to 100, step 10) were tested. For each combination, the sets of paired SNs with JSI=1 and with the specific threshold requirements were identified, and for each set the difference between the dates in the paired SNs was calculated, resulting in two temporal differences sets, $\{\Delta t_{JSI=1}\}$ and $\{\Delta t_{JSI<1}\}$. The best threshold combination was then obtained by applying to them the Baumgartner-Weiss-Schindler (BWS) test [35] and selecting the minimum threshold values for which the BWS test returned a p-value ≥ 0.01, so that the null hypothesis (i.e., $\{\Delta t_{JSI=1}\}$ and $\{\Delta t_{JSI<1}\}$ are generated from the same population) could not be rejected. The EMDN codes of these potentially equal SNs were compared with the ER results, as another measure of model performance.

IV. Results

The results of the fine-tuned NER model applied to the test set are reported in Table II. The F1-score for the class MD, 0.89, was slightly lower than that for ORG, 0.97. The micro-averaging, macro-averaging, and weighted-averaging F1 were all larger than 0.9, thus implying a good performance in general.

TABLE II. Performance of the fine-tuned NER model

	Precicion	**Recall**	**F1-score**
Medical device name (MD)	0.91	0.87	0.89
Organization name (ORG)	0.97	0.97	0.97
Micro-averaging F1	0.93	0.91	0.92
Macro-averaging F1	0.94	0.92	0.93
Weighted-averaging F1	0.93	0.91	0.92

TABLE III. Performance of the ER model on different EMDN code levels

	N	**EMDN code levels**			
		Level 1	***Level 2***	***Level 3***	***Level 4***
The same SNs	37	91.89%	91.89%	91.89%	78.38%
Potentially equal SNs	472	96.19%	94.70%	91.74%	83.05%

To evaluate the performance of the ER subtask, the first test was based on the comparison of the EMDN considering as gold standard those SNs with a JSI=1: as reported in Table III, it showed a good performance but due to the small number of samples (only 37 SNs), the results could appear as biased. So, a second test based on the potentially equal SNs, determined by the results of the BWS test using as thresholds 0.6, 0.9, and 60 for JSI, COS, and LD, respectively, was performed. Based on these conditions, 472 paired SNs were identified as potentially equal, and their EMDN codes were assumed as the gold standard to be compared with the ER results: high percentages of matched EMDN codes up to the fourth level were found, thus confirming the previously observed good performance. To note, a 100% ideal correspondence for EMDN codes is unrealistic due to the unsystematic nature of reporting systems, other than the multiplicity of languages.

V. CONCLUSION

The proposed approach extends a previously proposed framework [4] to allow aggregating publicly available SNs from national websites even when information is partially missing or not properly structured, as in the case of the Dutch SNs. The problem of missing lists of devices available on the market and relevant EMDN was also addressed by using available datasets from other countries. While extended to SNs from multiple countries, such a tool could facilitate PMS, assist EP in their evaluation procedure and provide useful information for national regulators.

ACKNOWLEDGMENT

We thank Professors Cinzia Cappiello and Maristella Matera, from Politecnico di Milano, Italy, for providing helpful advice throughout the work.

REFERENCES

[1] “Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC." *EUR-Lex.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2017.117.01.0001.01.ENG&hx0026;toc=OJ:L:2017:117:TOC

[2] J. Pane *et al.*, “EU postmarket surveillance plans for medical devices,” *Pharmacoepidemiology and Drug Safety*, vol. 28, no. 9, pp. 1155-1165, July 2019.

[3] A.G. Fraser *et al.*, “Improved clinical investigation and evaluation of high-risk medical devices: the rationale and objectives of CORE-MD (coordinating research and evidence for medical devices),” *EFFORT Open Reviews*, vol. 6, no. 10, pp. 839-849, Oct. 2021.

[4] Y. Ren, M. Bertoldi, A. G. Fraser, and E. G. Caiani, “Validation of CORE-MD PMS Support Tool: A Novel Strategy for Aggregating Information from Notices of Failures to Support Medical Devices’ Post-Market Surveillance,” *Therapeutic Innovation & Regulatory Science,* vol. 57, no. 3, pp. 589-602, Jan. 2023.

[5] DG Health and Food Safety, "The CND nomenclature: classificazione Nazionale dei Dispositivo medici," Jan. 2020. [Online]. Available: https://health.ec.europa.eu/system/files/2020-09/md_cnd_general_principles_en_0.pdf

[6] Health and Youth Care Inspectorate, “Safety notices.” [Online]. Available: https://english.igj.nl/medical-technology/safety-notices

[7] Italian Ministry of Health, “Open data.” [Online]. Available: https://www.dati.salute.gov.it/dati/

[8] National Authority of Medicines and Health Products, “Infarmed.” [Online]. Available: https://www.infarmed.pt/

[9] “Selenium,” *PyPI*. [Online]. Available: https://www.selenium.dev/

[10] “ChromeDriver,” *ChromeDriver – WebDriver for Chrome*. [Online]. Available: https://chromedriver.chromium.org/

[11] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," *The 16th International Conference on Computational Linguistics*, vol. 1, 1996.

[12] O. Etzioni *et al.*, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial intelligence,* vol. 165, no. 1, pp. 91-134, June 2005.

[13] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of Biomedical Informatics,* vol. 46, no. 6, pp. 1088-1098, Dec. 2013.

[14] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," *Computer Science Department Faculty Publication Series,* vol. 11, 2003.

[15] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," *arXiv preprint arXiv:1702.02098,* Feb. 2017.

[16] J. Li, A. Sun, J. Han and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering,* vol. 34, no. 1, pp. 50-70, Mar. 2020.

[17] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems,* pp. 30, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, Oct. 2018.

[19] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116,* Nov. 2019.

[20] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing,"in *Proc. of the 2020 conference on empirical methods in natural language processing: system demonstrations,* pp. 38-45, Oct. 2020.

[21] L. Tunstall, L. Von Werra, and T. Wolf, *Natural Language Processing with Transformers*. O'Reilly Media, Inc., 2022.

[22] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, Aug. 2018.

[23] E. F. Sang and J. Veenstra, "Representing text chunks," *arXiv preprint cs/9907006,* July 1999.

[24] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[25] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences,* vol. 12, no. 16, pp. 4102-4107, 2017.

[26] “Unidecode,” *PyPI.* [Online]. Available: https://pypi.org/project/Unidecode/

[27] “cleanco – clean organization names,” *GitHub*. [Online]. Available: https://github.com/psolin/cleanco

[28] L. Getoor and A. Machanavajjhala, "Entity resolution: theory, practice & open challenges," *Proceedings of the VLDB Endowment,* vol. 5, no. 12, pp. 2018-2019, Aug. 2012.

[29] H. Hapke, C. Howard and H. Lane, *Natural language processing in action: Understanding, analyzing, and generating text with Python*. Simon and Schuster, 2019.

[30] L. A. Zadeh, "Fuzzy logic," *Computer,* vol. 21, no. 4, pp. 83–93, Apr. 1988.

[31] “fuzzywuzzy,” *PyPI.* [Online]. Available: https://pypi.org/project/fuzzywuzzy/

[32] A. Koschmider, V. Torres and V. Pelechano, "Elucidating the mashup hype: Definition, challenges, methodical guide and tools for mashups," in *Proc. 2nd Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web at WWW,* pp. 1-9, Apr. 2009.

[33] E. Windmill, *Flutter in action*, Simon and Schuster, 2020.

[34] M. Levandowsky and D. J. Winter, "Distance between Sets," *Nature,* vol. 234, no. 5323, pp. 34–35, Nov. 1971.

[35] W. Baumgartner, P. WeiB and H. Schindler, "A nonparametric test for the general Two-Sample problem," *Biometrics,* vol. 54, no. 3, pp. 1129–1135, Sep. 1998